A new feature selection method based on stability theory - exploring parameters space to evaluate classification accuracy in neuroimaging data

Jane M Rondina[1,2], John Shawe-Taylor[2], and Janaina Mourao-Miranda[1,2]

[1] Centre for for Neuroimaging Sciences, Institute of Psychiatry, King's College London

[2] Department of Computer Science, Centre for Computational Statistics and Machine Learning, University College London

jane.rondina@kcl.ac.uk

Abstract. Recently we proposed a feature selection method based on stability theory. In the present work we present an evaluation of its performance in different contexts through a grid search performed in a subset of its parameters space. The main contributions of this work are: we show that the method can improve the classification accuracy in relation to the wholebrain in different functional datasets; we evaluate the parameters influence in the results, getting some insight in reasonable ranges of values; and we show that combinations of parameters that yield the best accuracies are stable (i.e., they have low rates of false positive selections).

1 Introduction

Feature selection (FS) methods applied to neuroimaging have increasingly become a discussion target. As several applications have been performed successfully using the whole brain (e.g. [1–5]), there have been some questions about the role of FS in neuroimaging applications using learning algorithms [6]. One of the multivariate approaches most commonly applied in classification based in neuroimaging is Recursive Feature Elimination (RFE) [7, 8], usually embedding Support Vector Machine (SVM) [9, 10]. However, this method has recently suffered some criticism. According to [11], since SVM results degrade with the increasing number of features, it is not clear whether the ranking provided by the initially trained classifier is a reliable measure for the elimination of voxels. Therefore, more stable approaches have been pursued, not only in order to increase accuracy in classification, but also as a strategy for mapping (enabling to localize features that best discriminate groups with sparsity based in stability instead of an arbitrary threshold).

Recently we presented a FS method - SCoRS (Survival Count on Random Subspaces) [12] based on a novel theory on Stability Selection [13]. In tat work, using a blocked functional dataset we showed that SCoRS improved the classification accuracy up to 10% using as few as 2.3% of the total number of features.

We also made a comparison with RFE and showed that SCoRS presented a better accuracy it was more stable (i.e, there were less false positive voxels) then RFE consistently in all the folds. In the present paper we investigate the effect of SCoRS parameters in different neuroimaging datasets.

2 Data and methods description

As neuroimaging comprehends a wide diversity of modalities, types of measurements and voxel's resolutions, in order to explore the effect of the parameters in different scenarios we used three real datasets. Their characteristics are described in the following table.

Table 1. Datasets description

Id	Dimensionality	Purpose
Dts1	36 X 27752	depressed patients *versus* healthy controls
Dts2	38 X 171601	depressed unipolar *versus* depressed bipolar
Dts3	42 X 140241	schizophrenic episodic *versus* continuous

SCoRS is based on iterative sub-sampling of features (subspaces) and application of a L1-norm regression (LASSO [14]) on them in order to select features which present non-zero coefficients more frequently. Considering that the subsampling is performed in a random way, the surviving features are expected to be stable under perturbation, as in each iteration the regression is applied to a different combination of variables. Its algorithm depends on three parameters: size of the subspaces, number of iterations and a final threshold (applied to eliminate features selected less frequently). In the present work we implemented a grid search to combine different values in discrete ranges defined through progressions fixed as: S (size of the subspaces), I (number of iterations) and T (threshold). Variables p and n represent the total number of features and the number of observations, respectively.

$$S = \frac{p}{2^i * n}, \text{ where } i = 4, 3, 2, 1, 0, -1, -2, -3, -4 \tag{1}$$

$$I = i * r, \text{where } i = 1:9 \text{ and } r \text{ was fixed as } 10^3 \tag{2}$$

$$T = i * r, \text{where } i = 1:9 \text{ and } r \text{ was fixed as } 10^{-1} \tag{3}$$

Each individual combination of parameters was performed inside a cross-validation, leaving out one subject per group in each fold. Afterwards, the subjects were classified using SVM.

One important issue related to stability of FS algorithms is how to quantify its susceptibility to variations in the training set. We implemented a false positive test in the following way:

I) Randomly choose 10% of the features selected;

II) Permute features chosen in step I among the examples (each feature independently);

III) Run the complete FS procedure again in the permuted data matrix;

IV) Compute the proportion of features in the permuted set which continue to be selected;

Following figure shows a representation of the algorithm:

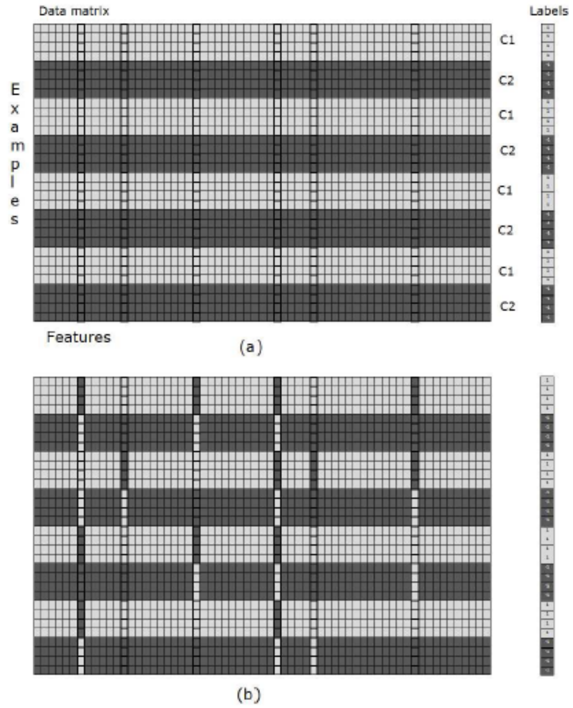

Fig. 1. Permutation algorithm representation

Ideally, none of the permuted features should be selected, as the permutation means to destroy the correlation between data and labels. However, if the number of examples is small, some partial correlation might still be kept as the number of possible permutations is limited.

Some recent studies have also applied the LASSo or Eastic nets in the context of fMRI analysis, but using different approaches of our. [18] proposed a LASSO extension adding a generalized ridge penalty ($l2$ norm of a weighted combination of the model parameters) to the LASSO regression model and showed that the resulting optimization problem can be efficiently minimized with existing LASSO solvers. [17] extended [18] and proposed GSR, a general approach for enabling properties beyond sparsity to be incorporated as an integral part of sparse model learning. [16] apply LASSO-PCR to study placebo analgesia. [15] discuss application of LASSO and Elasticnet for predictors selection using a multimodal dataset.

3 Results

Following figures show accuracies resulting from classification after feature selection obtained from each combination of parameters for Dts1, Dts2 and Dts3, respectively . Each figure has 3 rows and 9 columns, representing each parameter variation. The first row corresponds to the different subspace sizes, where axes y represents the number of iterations and axes x threshold levels. The second row corresponds to the different numbers of iterations, where axes y represents subspace sizes and axes x represents the threshold levels. The third row corresponds to the threshold levels, where axes y represents subspace sizes and axes x the numbers of iterations. Colors represent classification accuracy. Horizontal lines are placed in the colorbars indicating the wholebrain accuracy.

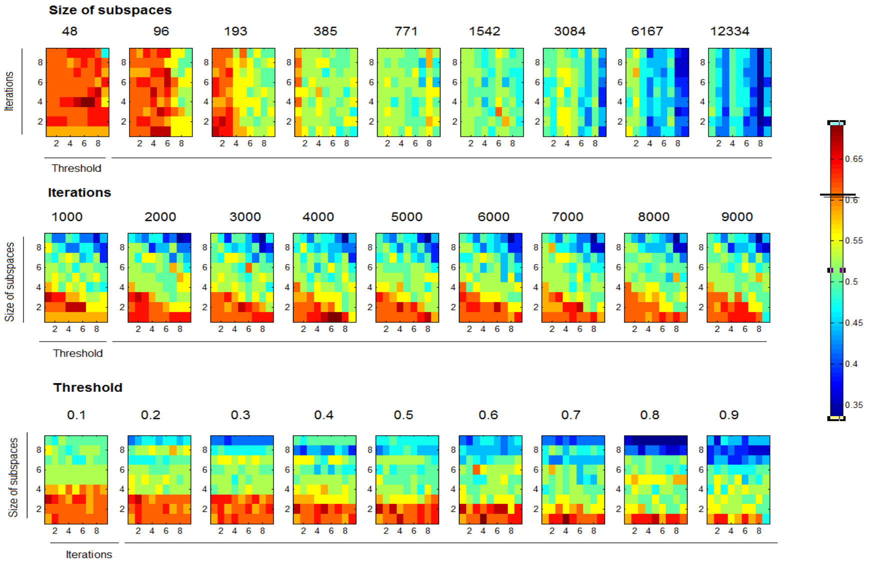

Fig. 2. Dataset 1

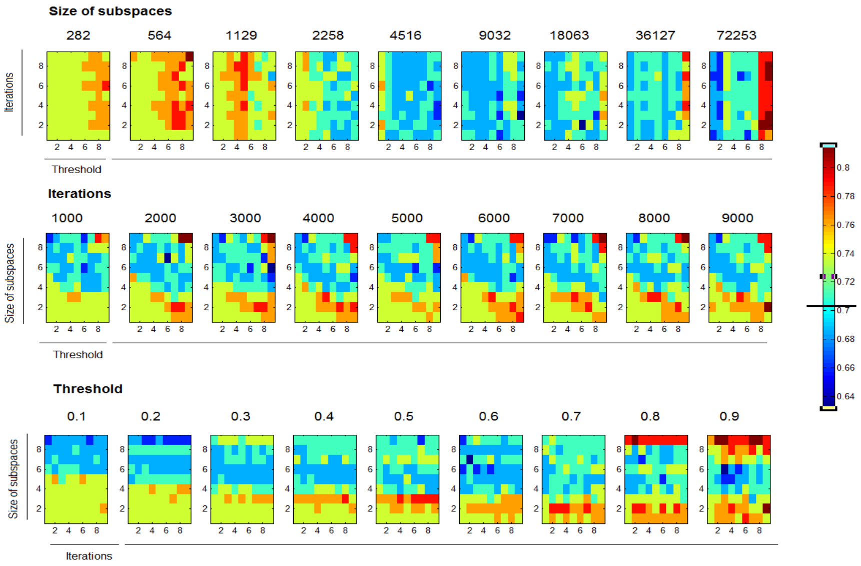

Fig. 3. Dataset 2

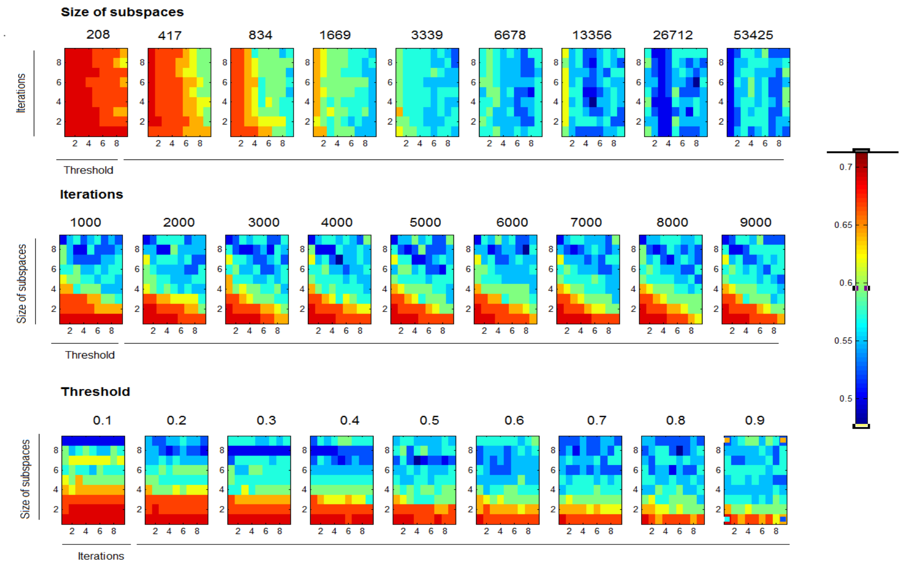

Fig. 4. Dataset 3

Figures 5, 6 and 7 show how many features are preserved with varying threshold in the same datasets (Dts1, Dst2 and Dst3, respectively). In each figure, nine graphs are presented, one for each size of subspace. In each graph, colored lines represent different numbers of iterations. Y axes in the graphs show the number of features selected. X axes show threshold levels (from 0.1 to 0.9).

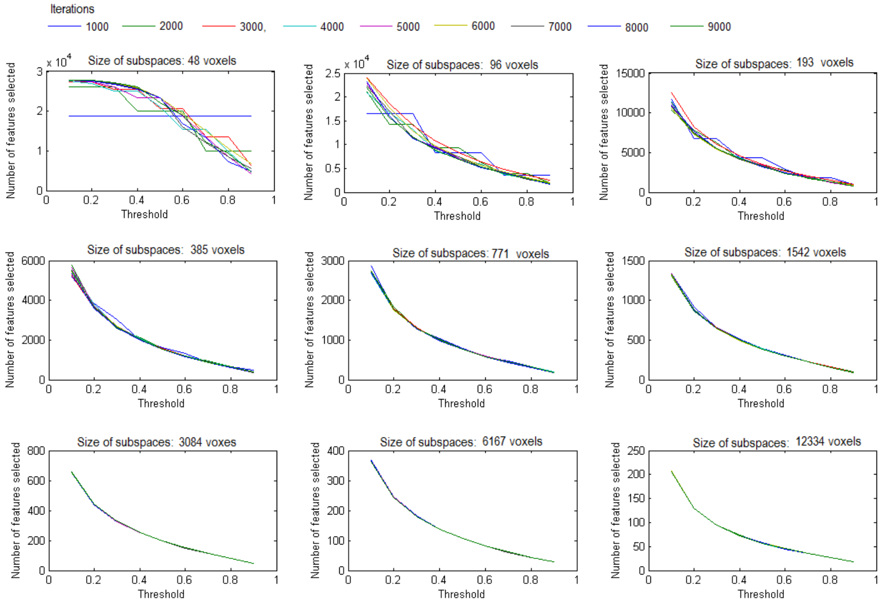

Fig. 5. Number of features preserved with varying threshold (dataset 1)

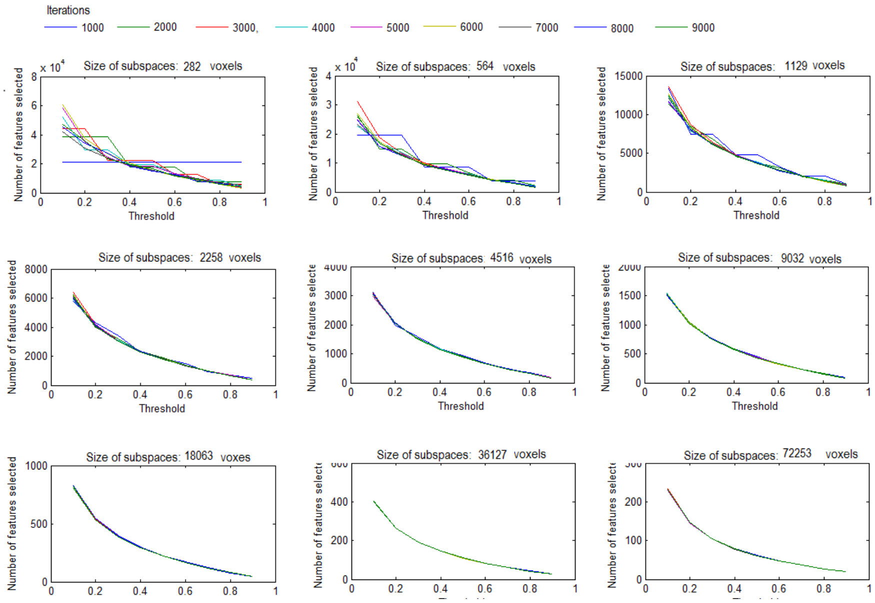

Fig. 6. Number of features preserved with varying threshold (dataset 2)

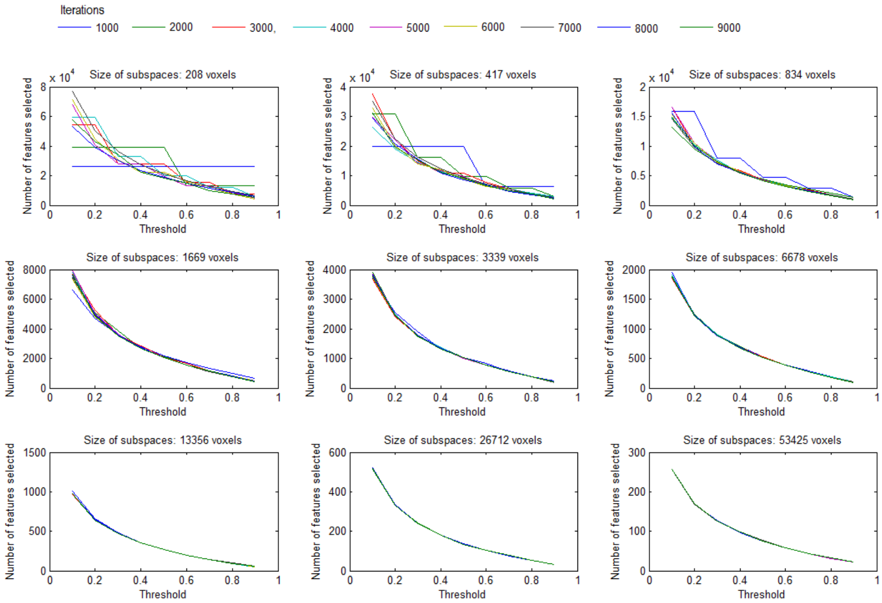

Fig. 7. Number of features preserved with varying threshold (dataset 3)

4 Discussion

Results have shown that the largest subspaces presented worse classification accuracy in all datasets. It makes sense, as for each iteration, the number of features selected by LASSO is limited to the number of examples. This causes results to be extremely sparse in case of large subspaces. The sparsity can also be verified in the graphs of figures 5, 6 and 7. In this case, the largest subspaces result in a number of features as small as less then 300 voxels (even for the minimum threshold) for all datasets.

The number of iterations did not have significant impact on the results, as can be easily seen in the second row of figures 2, 3 and 4. The same can also be corroborated in the figures 5, 6 and 7 where the graphs along theshold levels show very similar shape lines for different numbers of iterations. It is interessant to notice, however, that the higher the subspaces, the lines are closer together culminating in a complete overlap in the largest subspace.

The structural dataset did not show accuracy improvement in relation to the wholebrain. Other additional structural datasets have been tested also resulting in classification accuracy similar to the whole brain or around it (very slightly higher or lower). Additional investigations are necessary to understand the different behavior between structural and functional images. We hypothesize that this might occur because of the different nature of the measurements. Structural images are probability maps related to different tissues while functional images have absolute values related to oxygen levels. Other hypothesis is that the anatomical changes due to certain disorders might have different pattern of spreading, being more advantageous to use all the features in the most of the cases. Additional investigation is necessary for better understand the differences in FS performance between structural and functional images.

Interesting results were obtained through the false positive control (described in section 2.1). For the combination of parameters resulting in the highest accuracy, the false positive ratios for datasets 1 and 2 were respectively 0.0714 and 0.0557 (i.e. 7% and 5% of the selected voxels were false positive). This can be an encouraging indication towards the development of a multivariate mapping method with false positive control with potential to inferences, which would be of great appeal to clinical research. For this challenge, an approach able to control false negative would also be of great interest.

5 Acknowledgment

The authors thank Prof. Steve Williams (Dep. of Neuroimaging, Inst. of Psychiatry, KCL) for providing the first dataset, Prof. Mary L. Phillips (Dep. of Psychiatry, University of Pittsburgh, School of Medicine) for providing the second dataset and Dr. Paola Dazzan (Dep. of Psychosis Studies, Institute of Psychiatry, KCL) for providing the third dataset. JMR and JMM gratefully acknowledges support from the Wellcome Trust under grant no. WT086565/Z/08/Z.

References

1. Mourao-Miranda, J., Bokde, A., Born, C., Hampel, H., Stetter, M.: Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional mridata. Neuroimage. **95** (2005) 980-995.
2. Fu, C.H., Mourao-Miranda, J., Costafreda, S.G., Khanna, A., Marquand, A.F., Williams, S.C., Brammer, M.J.: Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. Biol Psychiatry. **63(7)** (2008) 656-62.
3. Koutsouleris, N., Meisenzahl, E., Davatzikos, C., Bottlender, R., Frodl, T. ad Scheuerecker, J. Schmitt, G., Zetzsche, T. ad Decker, P., Reiser, M., Mller, H., Gaser, C.: Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. Arch Gen Psychiatry **76** (2009) 700-712.
4. Marquand, A., M., H., Brammer, M., Chu, C., Coen, S., Mouro-Miranda, J.: Quantitative pre- diction of subjective pain intensity from whole-brain fmri data using gaussian processes. Neuroimage **49** (2010) 2178-2189.
5. Ecker, C., Rocha-Rego, V., Johnston, P., Mourao-Miranda, J., Marquand, A., Daly, EM., Brammer, M.J., Murphy, C., Murphy, D.G., MRC AIMS Consortium.: Investigating the predictive value of whole-brain structural mr scans in autism: a pattern classification approach. Neuroimage **49** (2010) 44-56.
6. Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehricy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O., Alzheimer's Disease Neuroimaging Initiative.: Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage **56(2)**(2010) 766-81.
7. Guyon, I., Elisseefi, A.: An introduction to variable and feature selection. Journal of Machine Learning Research **3** (2003) 1157-1182.
8. De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E.: Combining 19 multivariate voxel selection and support vector machines for mapping and classification of fmri spatial patterns. Neuroimage **43** (2008) 44-58.
9. Boser, B.E.; Guyon, I.M. and Vapnik, V.N.: A training algorithm for optimal margin classifiers. In D. Haussler, editor, 5th Annual ACM Workshop on COLT, Pittsburgh, PA, ACM Press (1992) 144152.
10. Vapnik, V.: Statistical Learning Theory. Wiley (1998)
11. Langs, G., Menze, B., Lashkari, D., Golland, P.: Detecting stable distributed patterns of brain activation using gini contrast. Neuroimage **56** (2011) 497-507.
12. Rondina, J.M., Marquand, A.F., Hahn, T.; Shawe-Taylor, J., Mourao-Miranda,J.: Selecting features based on stability to classify depressed patients in fMRI 17th Annual Meeting of the Organization for Human Brain Mapping, Quebec City, abstract 4181 (2011)
13. Meinshausen, N., Buhlmann, P.: Stability selection. Journal of the Royal Statistical Society. **72** (2010) 417-473.
14. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society **58** (1996) 267-288.
15. Bunea, F., She, Y., Ombao, H., Gongvatana, A., Devlin, K., Cohen, R.: Penalized least squares regression methods and applications to neuroimaging. Neuroimage **55(4)** (2011) 1519-1527.

16. Wager, T.D., Atlas, L.Y., Leotti, L.A., Rilling, J.K.: Predicting individual differences in placebo analgesia: contributions of brain activity during anticipation and pain experience. Journal of Neuroscience **31(2)** (2011) 439-452.
17. Ng, B., Abugharbieh, R.: Generalized sparse regularization with application to fMRI brain decoding. Inf Process Med Imaging **22** (2011) 612-623.
18. Ng, B., Abugharbieh, R., Varoquaux, G., Poline, J.B., Thirion, B.: Med Image Comput Comput Assist Interv. **14(2)** (2011) 285-292.
19. Funamizu, A., Kanzaki, R., Takahashi, H.: Distributed representation of tone frequency in highly decodable spatio-temporal activity in the auditory cortex. Neural Netw **24** (2011) 321-322.
20. Fan, Y., Shen, D., Davatzikos, C.: Classification of structural images via high-dimensional image warping, robust feature extraction, and svm. Med Image Comput Assist Interv **8** (2005) 1-8.
21. Mourao-Miranda, J., Reinders, A. A. T. S., Rocha-Rego, V ., Lappin, J., Rondina, J., Morgan, C., Morgan, K. D., Fearon, P., Jones, P. B. , Doody, G. A., Murray, R. M., Kapur, S., Dazzan, P.: Individualised Prediction of Illness Course at the First Psychotic Episode: a Support Vector Machine MRI Study. Psychological Medicine (2011)